The TREC-6 Spoken Document Retrieval Track

Ellen Voorhees, John Garofolo
National Institute of Standards and Technology
Gaithersburg, MD 20899

Karen Sparck Jones
Cambridge University
Cambridge CB2 3QG, U.K.

ABSTRACT

The Text REtrieval Conference (TREC) workshops provide a forum for different groups to compare retrieval systems on common retrieval tasks. The 1997 TREC workshop will feature a Spoken Document Retrieval task for the first time. This paper motivates the task and describes the measures to be used to evaluate the effectiveness of the retrieval methodologies.

1. The Text REtrieval Conference

The Text REtrieval Conference (TREC) series is cosponsored by the National Institute of Standards and Technology (NIST) and the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA) as part of the TIPSTER Text Program. The series, which started in 1992, is designed to promote research in information retrieval by providing appropriate test collections, uniform scoring procedures, and a forum for organizations interested in comparing their results. Thirty-eight groups including representatives from nine different countries participated in TREC-5 in November, 1996.

TREC has two main tasks, ad hoc and routing retrieval. The ad hoc task investigates the performance of systems that search a static set of documents using novel queries; the routing task investigates the performance of systems that use standing queries to search new streams of documents. In addition, TREC has smaller "tracks" that allow participants to focus on particular subproblems of the retrieval task. Recent track tasks have included Spanish retrieval, Chinese retrieval, the use of natural language processing techniques for retrieval, and retrieval of documents that result from paper documents being scanned by an Optical Character Recognition (OCR) process.

The retrieval of OCR documents was the focus of the TREC-5 "Confusion" track. The Confusion Track investigated methods for retrieving document surrogates whose true content has been confused or corrupted in some way. A different form of corruption will be used in TREC-6: retrieving spoken documents (i.e., recordings of speech) through surrogates produced by speech recognition systems. This new track, the Spoken Document Retrieval (SDR) track, is intended to foster research on retrieval methodologies for spoken documents. A second goal of the track is to encourage collaboration between the speech and retrieval research communities.

This paper defines the particular task to be addressed in the SDR Track and motivates the track's design. A detailed specification of the track, including sign-up procedures, samples of the data formats, and particulars of result submission, can be found at

`http://www.itl.nist.gov/div894/894.01/sdr97.txt` .

More information about TREC itself can be found at

`http://www-nlpir.nist.gov/trec` .

Questions about the track can be sent to either (or both) of the track organizers at `john.garofolo@nist.gov` or `ellen.voorhees@nist.gov`.

2. The SDR Track

The SDR Track was designed to encourage as much participation as possible in keeping with TREC's retrieval charter. The track therefore offers two modes of participation: SDR for those with speech recognizers and Q(uasi)SDR for those without. The latter is intended as a startup for those in the present retrieval community without immediate access to speech processing expertise[1]. While offering both options limits the experimental comparisons that can be made among groups and complicates the track definition, we anticipate that it will greatly expand the number of retrieval methodologies represented in the track.

2.1. Documents

The track will use stories (i.e., documents) taken from the Linguistic Data Consortium (LDC) 1996 Broadcast News corpus. This data was used in the November 1996 "Hub-4" DARPA Speech Recognition Evaluation [1, 2]. The test set will consist of about 1000 stories representing 50 hours of recorded material. A story is generally defined as a continuous stretch of news material with the same content or theme (e.g. tornado in the Caribbean, fraud scandal at Megabank), which will have been established by hand segmentation of the news programs. Note, however, that some stories such as news summaries may contain topically varying material, and that a story is likely to involve more than one speaker, include background music or noise, etc.

There will be four forms of the story data supplied for the track as shown in Figure 1 and described below.

SPH Sphere formatted speech files: digitized recordings of the broadcasts.

DTT Detailed TREC Transcriptions: hand-generated transcriptions used in speech recognizer training

[1] As this is a TREC track, all participants are required to produce retrieval output. Those in the speech community who do not have their own retrieval system may use a commercial retrieval system or a publically available system such as NIST's ZPRISE system.

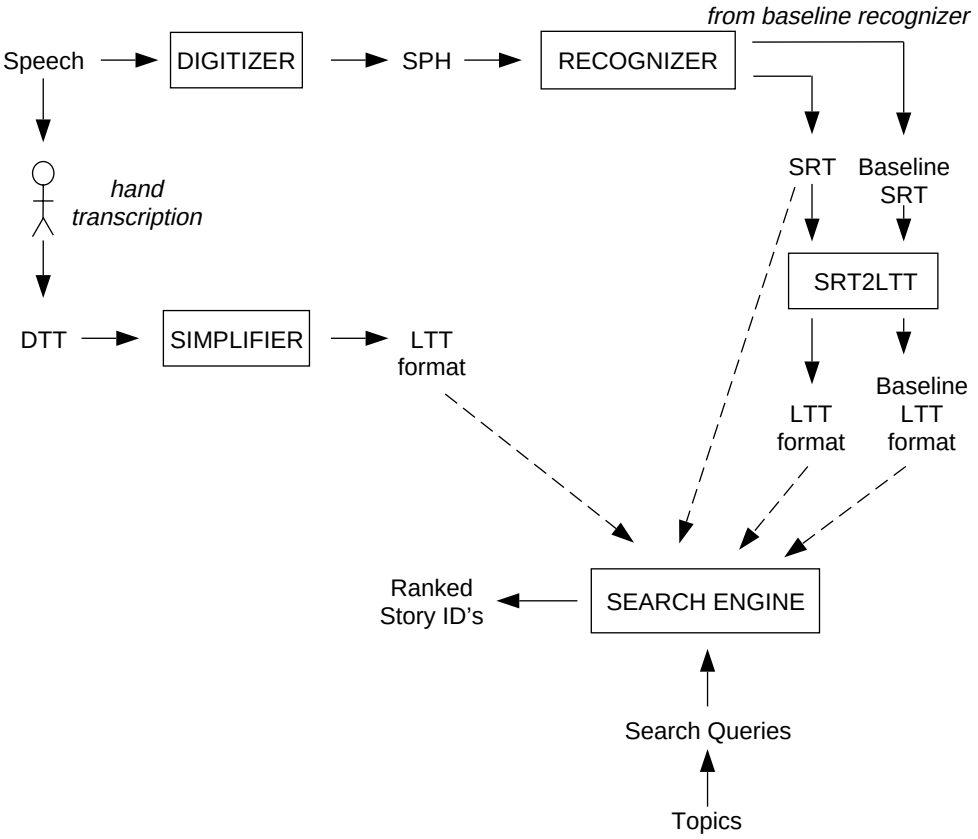

Figure 1: Data Flow in the SDR Track

and for speech recognition scoring that use the established DARPA broadcast news SGML-tagged annotation/transcription convention — hence not conveniently readable as simple text. (These are the LDC-generated transcripts with absolute Section (story) ID's added.)

LTT Lexical TREC Transcriptions: DTT's with most SGML tags removed and hence conveniently readable as text.

SRT Speech Recognizer Transcriptions: automatically generated transcriptions (assumed to contain recognition errors) produced by a particular recognizer when applied to SPH files. The file format is identical to LTT except that each word is bracketed by an SGML tag pair that indicates the time at which the word occurs. SRT's generated by a generous volunteer speech site will be used for Baseline testing (these are called the *Baseline SRT's*). A filter, "SRT2LTT", will be provided by NIST to strip the time tags to convert these to LTT format.

All of the above file types will be cross-linked by SGML-tagged time markers for story beginnings and ends. In addition, the following auxiliary files will be provided:

SIL Speaker Information Log: used to cross reference information about the speakers in the transcripts. This will be used primarily by speech sites in calibration.

NDX Index containing only Episode and Section tags used by the speech recognition systems to produce SGML-tagged SRT output for the evaluation.

2.2. The Task

A particular type of retrieval problem, called *known-item searching*, will be used in the SDR track. A known-item search is a retrieval task that simulates a user seeking a particular, partially-remembered document in the collection. In contrast to a more standard retrieval search where the goal is to retrieve/rank the entire set of documents that pertain to a particular subject of interest, the goal in the known-item search is to retrieve one particular document.

Known item searches were successfully used in the TREC-5 Confusion Track. Indeed, the searches are well-suited to this problem. When document content is corrupted, be it by OCR or speech processing, low-frequency words such as proper nouns and technical terms are the most affected. Yet low-frequency words are high-content-bearing words, and are precisely the words likely to be used to locate a specific document. Thus known-item searches exercise the parts of the retrieval methodologies that the track is most interested in. As a bonus, the searches do not require relevance assessments.

- Use of solar power by the Florida energy office.
- Excessive mark up of zero coupon treasury bonds.
- I am looking for a document about the dismissal of a lawsuit involving Adventist Health Systems.
- I am looking for theft data on the Chevrolet Corsica.
- efforts to establish cooperative breeding programs for the yellow crowned amazon parrot.
- morphological similarities between different populations of saltwater crocodiles.

Figure 2: Example Known-item Topics from the TREC-5 Confusion Track

Clearly, this obviates the need for relevance assessor's time — a critical resource at NIST. But it also means the track can run with fewer participants: since TREC uses pooled results to approximate exhaustive relevance assessments, the quality of the relevance assessments depends on the diversity of the pool and hence on the number of participants.

The search topics[2] will be produced at NIST and will be designed such that the author believes there is exactly one document in the collection that matches the topic. We expect there to be 50 topics created for the track. Example topics taken from the TREC-5 Confusion Track are given in Figure 2. Participants must use written forms of the topics for the required track runs. However, experiments with participants' own spoken versions are also welcome.

The set of retrieval runs for which results are to be submitted is given below. A retrieval run consists of running search queries for each of the 50 topics against a particular document set (see Figure 1). The set of runs required by the track were selected both to capture retrieval performance and to allow comparison between and within the SDR and QSDR Groups. We hope to gain insight into not only the overall performance levels obtainable, but also into how the speech recognition strategy and the retrieval strategy individually contribute to retrieval performance. The required retrieval runs are:

Speech (S1): a full SDR run with a site's own recognizer on the SPHERE-formatted digitized broadcast news recordings. (For SDR Group participants only)

Baseline (B1): a retrieval run using the the Baseline Speech Recognizer Transcriptions as input. This is the "speech run" for QSDR participants. However, SDR Group participants are also required to do this run to enable speech-based retrieval comparisons for all track participants.

Reference (R1): a retrieval run using the reference (hand-transcribed) Lexical TREC Transcriptions as input. This run enables retrieval-based comparisons across all track participants.

Participants may optionally submit a second Speech run and a second Baseline run to test the effects of variations in their own system parameter settings.

[2]Statements of information need are called "topics" in TREC to distinguish them from "queries" that actually get submitted to retrieval systems.

These required runs support retrieval performance comparisons as follows:

- between members of the SDR Group (i.e. the 'real' spoken document retrieval case), as a black box comparison not distinguishing the relative contributions of the recognition strategy and the retrieval strategy.
- between members of the QSDR Group, to compare retrieval strategies for the one shared recognition strategy, i.e. the one that delivered the Baseline Speech Recognizer Transcriptions.
- between all participants, SDR and QSDR, to compare retrieval strategies, via the Baseline Speech Recognizer Transcriptions.
- for each participant, between spoken document retrieval and text retrieval using the Lexical TREC Transcriptions, to calibrate the former against the latter for the participant's own retrieval strategy.
- for all participants, on text retrieval with the Lexical TREC Transcriptions, to compare retrieval strategies.

Together these runs permit a variety of comparisons to be made. The Lexical TREC Transcription text runs demonstrate what the level of performance would be for the given documents and topics with a perfect speech recognizer and the teams' various retrieval strategies. On the other hand, the baseline and individual recognizer runs demonstrate the effects the various recognizers have on retrieval performance.

2.3. Evaluation

Despite the fact that the track is using known-item searches, participants will be required to submit a full ranking of the collection (ordered in decreasing likelihood of the document being the known item) for each topic for evaluation. Experience has shown that a measure of simply success/fail for the first document retrieved is too stringent for both plausible topics and the realities of speech or retrieval systems. In addition, participants in the SDR Group will also submit recognizer output that will be evaluated using the traditional DARPA/NIST CSR word-error based metrics [3].

Since the traditional retrieval effectiveness measures of recall and precision are uninformative for known-item searches, other measures must be used. We investigated three different measures in the TREC-5 Confusion Track [4], and these measures will be used in the SDR Track as well. Other measures may also be introduced in the SDR track if further research

produces more appropriate measures. In all cases, the measures are based on the ranks assigned to the known items. A sample evaluation from the TREC-5 confusion track is shown in Tables 1 and 2.

The task in the TREC-5 Confusion Track was to rank the top 1000 documents per topic on each of three different versions of the *1994 Federal Register*: the correct copy, a scanned copy that had approximately a 5% character error rate, and a scanned copy that had approximately a 20% character error rate. These sets of documents correspond to the LTT and two different SRT transcripts in the SDR Track. The rank at which the known item was retrieved for each of the three versions for all 49 topics for the Confusion Track example is given in Table 1. A document that was not retrieved at all in the top 1000 documents was assigned a rank of 2000. Sites are asked to rank the entire collection in the SDR Track since this will preclude the need for an artificial "not retrieved" rank and thus eliminate discontinuities in the effectiveness measures.

The raw ranks are used to compute the measures given in Table 2. The first measure, "Histogram", counts the number of topics for which the known item was found in a certain range of ranks. Since the SDR Track will not have a "Not found" category (the full collection is ranked), we will use the ranges of 1–5, 1–10, 1–20, and 1–100. The overlapping categories in the histogram permit the histogram counts to be compared across sytems (system A may have fewer documents found in ranks 6–10 than system B because it has *more* documents found in ranks 1–5). The histogram counts are then equivalent to precision after 5 documents retrieved, after 10 documents retrieved, etc., which are common measures used in the rest of TREC.

The second measure, labeled "Mean rank when found", is the mean rank at which the known item was found averaged across all topics that retrieved the known item in the top 1000 documents. This measure gives an easily-interpreted idea of how well the retrieval methodology ranks the known item if it finds it at all. Since the SDR track will rank all documents, the average will always be computed over all 50 topics. (When the average is computed over all topics, this measure is also known as *expected run length*.)

The last measure is called the "Mean reciprocal rank". It is the mean of the reciprocal of the rank at which the known item was found over all the topics, using 0 (not 1/2000) as the reciprocal for topics that did not retrieve the known document. Unlike the mean rank when found measure, this measure penalizes runs that did not retrieve a known item while minimizing the difference between, say, retrieving a known item at rank 750 and retrieving it at rank 900. It is also bounded between 1 and 0, inclusive, so the measure is interpretable without knowing how many documents were ranked. Indeed, since there is only one relevant document per query, the reciprocal rank of that document is the precision at that document, and therefore it is the average precision of the query as well (average precision is the precision averaged over all relevant documents of the query). Average precision is another frequently used measure in the other parts of TREC, so "mean reciprocal rank" gives some basis of comparison with other retrieval methods.

3. Conclusion

The Spoken Document Retrieval Track is intended to foster research on indexing and retrieving spoken documents. While the SDR problem has parallels to the problem of retrieving documents that have been corrupted due to OCR errors, solutions to the two problems are likely to be quite different since the nature of the corruption differs in the two cases. Whereas OCR errors tend to turn words into non-words, speech recognition errors tend to substitute other actual words for correct words.

The TREC-6 SDR Track is the initial offering of a spoken document retrieval track and as such must be viewed as something of an experiment itself. The results of the track are sure to be preliminary if only because a 1000-document collection — a formidable challenge to produce from 50 hours of speech — is very small for retrieval experiments. But we strongly encourage active participation in this track in order to gain sufficient experience with the SDR problem to guide future research.

References

1. Garofolo, J.S., Fiscus, J.G., and Fisher, W.M., "Design and preparation of the 1996 HUB-4 broadcast news benchmark test corpora," *Proceedings of the DARPA 1997 Speech Recognition Workshop*, 1997.
2. Graff, D., Wu, Z., MacIntyre, R., and Liberman, M., "The 1996 broadcast news speech and language-model corpus", *Proceedings of the DARPA 1997 Speech Recognition Workshop*, 1997.
3. Pallett, D.S., et al., "1996 Preliminary broadcast news benchmark tests", *Proceedings of the DARPA 1997 Speech Recognition Workshop*, 1997.
4. Kantor, P., and Voorhees, E.M., "The TREC-5 confusion track", *The Fifth Text REtrieval Conference (TREC-5)*, to appear.

	Correct	5%	20%
1	1	1	2
2	8	15	44
3	2	2	1
4	5	11	24
5	1	1	1
6	1	1	1
7	2	2	2
8	1	1	1
9	2	6	2
10	1	1	8
11	1	1	2
12	1	125	92
13	3	3	3
14	2	1	97
15	1	462	2000
16	1	1	1
17	1	7	11
18	1	2	10
19	1	1	93
20	1	1	9
21	2	2	9
22	1	1	3
23	1	1	2
24	6	7	4
25	1	1	1
26	1	1	2
27	1	1	18
28	6	13	384
30	16	39	60
31	2	2	3
32	7	10	29
33	1	1	1
34	1	3	23
35	1	1	1
36	1	94	981
37	1	1	37
38	1	9	23
39	1	15	342
40	26	138	435
41	1	1	5
42	1	1	5
43	1	13	14
44	1	1	1
45	1	1	1
46	11	103	2000
47	266	119	186
48	2	3	425
49	1	1	6
50	5	2	156

Table 1: Raw Ranks of an Example TREC-5 Confusion Track Submission

Histogram			
Number of items found at rank r where			
	Correct	5%	20%
$1 \leq r \leq 10$	45	37	27
$10 < r \leq 100$	3	7	13
$100 < r \leq 1000$	1	5	7
Not found	0	0	2

	Correct	5%	20%
Mean rank when found:	8.24	25.10	75.77
Mean reciprocal rank:	0.7506	0.5857	0.3285

Table 2: Evaluation Measures Computed for the Example TREC-5 Confusion Track Submission